Baseline testing: science or fantasy?

There's nothing hidden in your head
The Sorting Hat can't see,
So try me on and I will tell you
Where you ought to be.

The selection of children into houses at Hogwarts famously involves a magic 'sorting hat'. A fiction, of course, unlike baseline tests in real schools. The Government's baseline tests at the start of Reception produce numerical data, so they have an aura of scientific accuracy. **They are anything but.**

This article will focus particularly on the tests designed by CEM as one of the three approved providers of Reception Baseline Assessment in September 2015. This is not because CEM are incompetent but rather the opposite: they were the most experienced providers. Their test was based on PIPS, sold commercially to hundreds of schools in various countries and refined over more than a decade.

Predictive validity

CEM marketed its baseline tests as having 'excellent predictive validity'. Our subsequent investigations showed this to be a dubious claim. Perhaps their advertising department really meant 'This is about as good as it gets!'

Other CEM documents showed a correlation of around 0.7 between the PIPS test and attainment two years later. 0.7 is generally reckoned to be quite a strong correlation in the social sciences, but we have to ask questions of context and purpose. As one statistician in the Reclaiming Schools network - a former civil engineer - pointed out, when you're calibrating measuring instruments a correlation of 0.99 is disastrous: *bridges collapse*! A test with a correlation of 0.7 purporting to predict cardiac arrest or alzheimers in the next two years would be unusable due to far too many false negatives and false positives.

Around the same time, another colleague with many years of experience in educational statistics pointed out the need to square a correlation in order to estimate how much of an outcome can be predicted from an input measure. (This is because the formula for correlations involves a square root, which has to be cancelled by squaring.) Squaring 0.7 results in 0.49, i.e. around half. In other words, only half of the outcome (eg a child's KS1 result) is predictable on the basis of the baseline test.

Further data came to light following a Freedom of Information request for a 'chances table'. A chances table shows what proportion of children with each specific baseline score go on to reach different KS1 levels. It provides much more detailed information than a generalised correlation figure. The PIPS test made sound predictions of a KS1 sub-level for roughly 4 children out of 10. (England no longer use levels or sub-levels of course, but that was the basis at the time and the problem of predictive validity remains the same whatever symbols are used.)

The data worked quite well at the extremes: in particular, children with unusually high initial scores at PIPS tended to still do very well at KS1. It was rather more problematic for low scorers, many of whom reached average attainment just two years later. For the vast majority however baseline scores were a poor guide. As an example, of children with a baseline score of 50 out of 100, 6% were graded W or 1 at KS1, 13% received 2C, 28% 2B, 32% 2A, and 21% ended up with level 3. (These details are for KS1 Reading, from a baseline test at the end rather than the start of Reception.)

Readers will remember that the Government's intention was, and is, to link a baseline test undertaken in the first half term of Reception with KS2 results nearly seven years later. The DfE resolutely refused to factor in the child's month of birth, a major factor for such young children. Tests had to be in English even for children speaking another language at home. CEM's own researchers have also raised serious doubts about the predictive capacity of these tests in terms of emerging special educational needs.

Half a lesson learnt

As explained earlier, the above analysis relates to an organisation with substantial expertise in predictive testing. CEM have an established reputation based on a bank of tests for different age groups. Judging by the data we saw from one of the other providers, it is shocking how lax the DfE were in vetting these providers. It seems that neither of the other two approved providers had any longitudinal data to underpin their bid.

The DfE soon realised they were presiding over a disaster and commissioned an independent evaluation from the Scottish Qualifications Authority. Unfortunately the DfE's official conclusion was simply that data from the different providers could not be aligned. DfE officials failed to acknowledge - and perhaps did not even question - the reliability of any single provider.

Testing very young children is particularly fraught with difficulties. A major problem is that test items are often borrowed from tests for older children. In other words, items originally

designed to check whether a 7-year-old *has learnt* how to do something are used to determine whether a 4-year-old *will be able to learn it*. Absurdly inappropriate (and to many children, meaningless) test items were used such as:

> Say the word *parrot* without the P.
>
> I am going to sound out a word like a robot would say it: *p-i-n*. Can you tell me what word I have sounded out?

or the criterion:

> Can the child order and ascribe numbers up to 20 and add and subtract using single digit numbers?

Dividing up words into separate sounds is an artificial exercise which accompanies early literacy teaching: children are not born with the ability to divide meaningful words into phonemes. Whether a four-year-old can already manipulate numbers 1-20 reflects both test-preparation received from parents or nursery, and the child's general level of maturation. To assume that such test items are reliably predictive of subsequent attainment is delusional.

Selection at four

Unfortunately such predictive assessment could work, to an extent, as self-fulfilling prophecy. In other words, if you label all the low-scorers as "Low Ability" and treat them as such, and teach the high-scorers on a "High Ability" table, there is a stronger chance that they will reach the predicted attainment levels.

Sadly such practices - illegal and unthinkable in Scandinavia - have become all too common in today's test-driven primary schools. Some heads clearly believed the baseline tests would enable them to determine each child's *ability* and *potential* in the first few weeks of Reception.

It can be argued, of course, that baseline tests are only intended as a measure at whole-school level i.e. that by comparing the whole school's aggregate baseline scores with its aggregate KS2 scores, you can judge the school's "effectiveness". Regardless of the assumptions behind such judgements, there is a clear problem: you can't do this without putting a label on each individual child.

Numerical data: the aura of science

Others have suggested that the Early Years Foundation Stage Profile should be used instead. The Profile certainly a more holistic register of development than baseline tests focused on (proto-) literacy and numeracy, but once its descriptions are converted into numerical data,

we encounter similar problems to formal baseline tests. The DfE tried this several years ago (DfE Research Report RR034). Of children with the midpoint score on reading in the EYFS profile, 23% went on to get W or L1 at KS1 Reading, 22% got 2C, 31% got 2B, 18% got 2A, and 7% L3. This is about as useful as judging children by the cleanliness of their fingernails and shoes.

All of this forms part of a spurious search for certainty and the illusion that schools can be judged fairly by comparing value-added data. It forgets that, in general terms, children in poverty *tend* towards less progress even with talented teachers. It is blind to the potential advantages that accrue from university-educated parents. It overlooks the difficulties of predicting progress for EAL children with variable exposure to English. It neglects the levels of mobility in and out of many urban schools. It ignores the variation between one year group and the next, exacerbated in smaller schools. It forgets that one sick child and an acrimonious divorce can upset the aggregate score of a one-form-entry school.

Unfortunately the subtleties of all this have gone unnoticed by the largest headteachers' body. Clearly regarding a challenge to KS2 tests as inconceivable, the NAHT appear to have opted for Baseline as a fair basis for value-added comparisons. It is sad that such a respected organisation should accept a high-stakes accountability system designed to put around 10% of their members in front of a firing squad each year, and identify their schools for academisation (or increasingly, re-brokering into different chains). Is this evidence-based, or are there interests at work? A scrutiny of the NAHT's *Assessment Review Group* Report on KS1 and KS2 statutory assessment 'Redressing the balance' shows that Robert Coe of CEM was a full member, and Jan Dubiel of Early Excellence was a key contributor. Schools Minister Nick Gibb, when asked to provide evidence for baseline testing, alternately cites the NAHT and Professor Coe. There is a circularity when key researchers are simultaneously policy advisers to government and recipients of major contracts.

Hopefully, the parlous condition of government after the General Election on 8 June could make them rather more hesitant to relaunch baseline tests, provided there is sufficient pressure from experienced primary school teachers and heads.

Dr Terry Wrigley is Visiting Professor at Northumbria University, editor of the international journal Improving Schools, and one of the coordinators of the Reclaiming Schools research network. Further information can be found at www.reclaimingschools.org